PLISKA
STUDIA MATHEMATICA BULGARICA
ПЛИСКА
БЪЛГАРСКИ МАТЕМАТИЧЕСКИ СТУДИИ

For further information on
Pliska Studia Mathematica Bulgarica
visit the website of the journal http://www.math.bas.bg/~pliska/
or contact: Editorial Office
Pliska Studia Mathematica Bulgarica
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Telephone: (+359-2)9792818, FAX:(+359-2)971-36-49
e-mail: pliska@math.bas.bg

Pliska Stud. Math. Bulgar. **13** (2000), 5-14

PLISKA
STUDIA MATHEMATICA
BULGARICA

ALGORITHMIC METHODS IN QUEUES AND IN THE EXPLORATION OF POINT PROCESSES

Marcel F. Neuts

This is a review of methodology for the algorithmic study of some useful models in point process and queueing theory, as discussed in three lectures at the Summer Institute at Sozopol, Bulgaria. We provide references to sources where the extensive details of this work are found. For future investigation, some open problems and new methodological approaches are proposed.

Keywords: Markovian arrival processes, queues, matrix-analytic methods, algorithmic probability.

1 Introduction

Those experienced in probabilistic modelling are well aware of the stringent, often unrealistic, assumptions that are imposed to retain explicit computability of interesting descriptors of a model. They share the resulting frustration with workers in statistics, physical applied mathematics, and mathematical biology. The advent of modern computers radically changed the range of explicit computability. The rare explicit analytic solutions are now relatively less important. Although there is residual reluctance to recognizing that reality and to adapting mathematical education accordingly, the computer and algorithmic thinking are now indispensable to applied mathematicians.

Computational science: Between the purely theoretical exploration of models by mathematical formalism and the empirical, experimental study of prototypes, there lies the rich, still emerging methodology of computational science. In brief, that methodology seeks understanding of the qualitative behavior of the model though systematic numerical computation or simulation. For the computational scientist, the algorithm fulfills the same function as the laboratory tool for the experimental physicist. The mathematical rigor that the former puts into the construction of the algorithm is on a par with the depth of theoretical physical understanding that the latter must bring to experimentation. With the ongoing rapid evolution of computers and greater familiarity with their use,

computational science will merge, I am convinced, with all analysis and experimentation. It is, in fact, a natural bridge between these two approaches. They should never have been thought of as distinct to begin with.

Since the early 1970s, I have devoted my efforts to developing such methodology for probability models, and specifically, for queues and point processes.

All mathematical thought deals with idealized structures. When we study, say, a queueing model, we postulate an idealized version of the service mechanism and of the arrival process. To be appealing, that idealized version ought to: be mathematically sufficiently simple for analysis (tractability), be capable of representing a wide variety of qualitatively different features (versatility), and should lead to results that, with reasonable effort, can be implemented on today's computers (computability.) For example, the homogeneous Poisson process, that simplest of point processes, excels in tractability, has little or no versatility, and few results derived under Poisson assumptions offer a computational challenge. Sometimes, I illustrate that by saying that, among point processes, the Poisson process occupies a comparably privileged place as the constant function holds among functions of a real variable.

The first lecture dealt with versatile, algorithmically tractable generalizations of the exponential distribution and Poisson process, to wit, distributions of phase type (*PH*-distributions) and the Markovian arrival process (*MAP*).

When the Poisson input of classical queueing models, such as *M/G/1* queue, is replaced by a *MAP*, we can still use embedded Markov renewal processes, the traditional tools for their analysis. The transition probability matrices of these embedded processes retain the same formal structure as in the elementary case, but their elements are now matrices themselves. The preservation of *structure*, not of analytic detail, enables us to treat these generalized models in substantially the same way as the elementary case. Of course, along the way many mathematical results must be generalized to serve in this new setting. These developments, known collectively as the *matrix- analytic methods* for probability models, were reviewed in the second lecture.

The third lecture was devoted to thought experiments in which random transformations are applied to point processes to bring out quantifiable descriptors of their behavior. In relation to these, there are many open questions of a theoretical nature; there is room for methodological development in computer experimentation, data analysis and visualization.

2 PH-Distributions and the Markovian Arrival Process

The distributions of phase-type (*PH*-distributions) form a large class of probability distributions which is dense in the set of all distributions on the right half-line. Their importance lies in the tractable solutions they provide for many useful probability models. The Markovian arrival process (*MAP*) is a similarly tractable generalization of the Poisson process. Expositions of the basic properties of both are found in Lucantoni [**?**], and in Neuts [**?**] and [**?**]. Although it takes some practice to master the matrix formalism used in calculations, the methodology is elementary and constructive.

2.1 PH-Distributions

Any probability distribution that can be the probability law of the time to absorption in a finite Markov chain with a single absorbing state is *of phase type.* There are entirely similar developments for continuous and for lattice distributions of phase type. For the continuous case, let Q be the generator of a finite Markov chain with a single absorbing state, partitioned as

$$Q = \begin{pmatrix} T & \mathbf{T^0} \\ 0 & \mathbf{0} \end{pmatrix},$$

where T is a non-singular matrix, and $\mathbf{T^0}$ is a column vector such that $T\mathbf{e} + \mathbf{T^0} = \mathbf{0}$. Its initial probability vector is similarly partitioned as $(\boldsymbol{\alpha}, \alpha_*)$, where α_* is the probability of instantaneous absorption. The distribution of the time until absorption is then given by:

$$F(x) = 1 - \boldsymbol{\alpha} exp(Tx)\mathbf{e}, \text{ for } \mathbf{x} \geq \mathbf{0}, \tag{1}$$

and its density portion is

$$F'(x) = \boldsymbol{\alpha} exp(Tx)\mathbf{T^0}, \text{ for } \mathbf{x} > \mathbf{0}. \tag{2}$$

These are the generic forms of a *PH*-distribution and its density. Their similarity to the exponential distribution is obvious. The transient states of the Markov chain are called *phases.* The pair $(\boldsymbol{\alpha}, T)$ is called a *representation* of the *PH*-distribution. Representations are not unique. An interesting, difficult problem is to construct minimal representations of *PH*-distributions with either as few phases or as few parameters as possible.

The class of *PH*-distributions has many useful closure properties that allow operations such as convolution, mixing, or some forms of conditioning to be given matrix representations in terms of the representations of the components of these operations. The explicit construction of representations is of paramount importance. It often leads to large, but specially structured matrices. A very general closure theorem for *PH*- distributions is found in Assaf and Levikson [**?**], but at that level of generality the constructions of representations is not known, and is likely to be impossible.

Necessary and sufficient conditions for a probability distribution on $[0, \infty)$ to be of phase-type were established in a beautiful paper of O'Cinneide [**?**]. While any distribution on $[0, \infty)$ is the limit of sequences of *PH*-distributions, the actual construction of approximations is done by numerical procedures. An excellent discussion of these and of procedures for fitting *PH*-distributions to data, is found in the article by Asmussen, Nerman, and Olsson [**?**], which also reviews the related literature.

The *PH-renewal process*, the renewal process with an underlying phase-type distribution, leads to an elementary, but useful construction. Assuming, for ease of exposition, that $\alpha_* = 0$, we restart, upon each absorption, the absorbing Markov chain, instantaneously and independently of the past. Defining path functions by right-hand continuity,

we obtain an m- state Markov chain with generator

$$Q^* = T + \mathbf{T^0}\boldsymbol{\alpha}, \tag{3}$$

which, without loss of generality, can be made irreducible. The generator Q^* plays a crucial role in the derivation of explicit matrix formulas for the renewal function and the renewal density of *PH*-renewal processes. It also clearly shows that the *PH*-renewal process is a particular case of the *MAP*. It is now customary to develop the matrix formalism for *MAP*s and to obtain the results for the renewal process by noticing the simplifications induced by that fact that the matrix $D_1 = \mathbf{T^0}\boldsymbol{\alpha}$ is dyadic.

2.2 The MAP

In modelling, say, the arrivals to a queue, we are primarily interested in the interarrival times, the successive intervals between arrivals. Renewal processes and, most commonly the Poisson process, are the most familiar processes used in modelling. However, as we know, stochastic models requiring as few as two general renewal processes are, in all but a few cases, already analytically intractable. For example, queues whose input is the superposition of two independent non- Poisson, renewal streams defy analysis except in very special cases. To model a modest form of dependence in the input stream, the process of the successive intervals between transitions in a Markov renewal process, (*semi-Markovian arrivals*), has been considered, but the tractability of the corresponding models is severely limited.

The *MAP*, of which we briefly discuss only the continuous-time version, is generated by the transitions of a particular class of Markov renewal processes related to irreducible continuous-parameter Markov chains.

We consider an irreducible $m \times m$ generator D with invariant probability vector $\boldsymbol{\pi}$. We write the matrix D as the sum of matrices $D_0,\ D_1$, where D_0 has negative diagonal elements and all its remaining elements and those of D_1 are nonnegative. Moreover, D_0 is nonsingular and $[-D_0]^{-1}$ is a nonnegative matrix.

The *MAP* is the point process generated by the transitions epochs of the m-state Markov renewal process with transition probability matrix

$$F(x) = \int_0^x exp(D_0 u)du D_1, \text{ for } x \geq 0. \tag{4}$$

The *MAP* is parametrized by the matrices D_0 and D_1. The homogeneous Poisson process of rate λ arises when $D_1 = -D_0 = \lambda$. The *MAP* is closely related to the Markov chain with generator D. A nice, intuitively appealing description of how the *MAP* is constructed by adding a Markovian labeling of transitions is given in Lucantoni [**?**]. For a survey giving many examples of *MAP*s, see Neuts [**?**]. To name only one important property, the superposition of two independent *MAP*s is also a *MAP*. That allows us, at least in some cases, to provide a full analysis of the effect of superpositions. A practical situation where that arises involves the decision whether or not to accept an additional job stream to a single server queue with spare capacity for service.

The *Markov-modulated Poisson process*, (*MMPP*), is the special type of *MAP* where the matrix D_1 is diagonal, say, with diagonal elements $\lambda_1, ..., \lambda_m$. It is the doubly stochastic Poisson process whose rate assumes one of m values depending on the state of an underlying Markov chain with generator D. Statistical procedures for fitting *MMPP*s are studied in Rydén [**?**] and [**?**], and in references therein. Statistical methodology for general *MAP*s is more difficult and deserves further attention.

This is not the place for a detailed exposition of the matrix formalism of *MAP*s. Let me discuss instead why they are such an appealing, potentially very useful class of point processes. In the first place, they are a versatile generalization of the Poisson process which, in its matrix formalism, preserves some of the analytic tractability of the elementary case. That is particularly evident from the elegant results obtained for classical queueing models in which Poisson arrivals are replaced by *MAP*s.

In simulation methodology, it is very easy to generate realizations of *MAP*s with a variety of initial conditions. For example, with one particular, readily computed, initial probability vector, one obtains the stationary version of the *MAP*. That can serve to eliminate some sources of initialization effects that commonly plague simulations.

Because of their relationship to finite Markov chains and their matrix-analytic formalism, *MAP*s and *PH-* distributions can occasionally be used to prove results that are presently intractable for general distributions. Noteworthy examples of that technique are found in Neuts and Takahashi [**?**] and Neuts [**?**]. In the first paper, it is shown that the steady-state distributions in a complex multi-server queue have asymptotically exponential or geometric tails. That result is likely to be valid under broader assumptions on the tail behavior of general service time distributions, but the appropriate methods to prove such results do not yet exist.

Similarly, it is sometimes possible to prove general results by passing through the techniques of *MAP*s or *PH-* distributions. If the end results, at least for continuous functionals, do not involve the explicit form of *MAP*s or *PH-* distributions, they hold in general. That is a general consequence of the unique extension theorem for continuous functions on a dense subset of a topological space. That method is described in some detail and is utilized in [**?**].

However, the greatest appeal of *MAP*s lies, I believe, in their use as *benchmarks* in statistical or data-analytic studies of point process data. Our toolbox of mathematical descriptors of physical properties of point processes is still very limited; it mostly consists of second-order moment formulas. There is much discussion, notably in the telecommunications community, of the physical characteristics of data streams that are most crucial to design and performance. The imprecise term *burstiness* is used to describe qualities of data streams not readily captured by elementary, tractable point process models.

There is an ongoing investigation of random transformations of point processes that can elucidate their behavior in a quantitative way. An early example of a thought experiment that is a random transformation of a point process led to the notion of *peakedness*. One imagines the point process to be the input to an infinite server queue with independent, identically distributed holding times. Equivalently, one associates i.i.d lifetimes with each event in the point process and one derives the steady-state distribution of the number of events *alive* at an arbitrary time epoch. The peakedness is defined by the

coefficient of variation of that distribution. It depends, of course, on the distribution of the holding times and is explicitly computable only for special distributions, among them the exponential, and the *MAP* as the point process of interest. The peakedness, which can be readily estimated from data, provides a limited, though useful descriptor of the! variability of a point process. However, as it is presently defined, it is known to measure only second-order properties of the process. For a thorough discussion of peakedness, see Eckberg [**?**].

In the same spirit, my associates and I have investigated various other random transformations. Among these are *local poissonification* [**?**], *selective marking* [**?**], *competitions for runs*, [**?**], and alternative associations of lifetimes (clocks) to the successive events of a point process, see [**?**] and [**?**]. For *MAP*s - and possibly *only* for *MAP*s - we can express various distributions and other quantities of the induced descriptors by matrix formulas that can be computationally implemented. The idea is to ascertain that, for *MAP*s with well-understood behavior, certain particular descriptors indeed reflect that behavior informatively. If so, that will induce confidence using that descriptor also in cases where a *clean* theoretical analysis is not feasible. That situation is entirely analogous to that for many common statistical procedures whi! ch can be fully justified only under normality and independence assumptions.

Much more needs to be done. Aside from matrix-analytic studies of the random transformations, it is necessary to develop algorithms for the computation of the various distributions for *MAP*s and to study these distributions for representative models. Major insight is gained from the visualization of these random transformations applied to simulated data. However, finding the proper, most informative, and accurate visualizations is not always easy. They require much trial-and-error on the part of the investigator. This truly computation-intensive line of investigation is, in my opinion, highly promising.

3 Structured Markov Chains

In elementary queueing theory, two models, the $M/G/1$ and the $GI/M/1$ queues, allow particularly tractable analyses. They are extensively discussed in all introductory texts on queueing theory. Because of the Poisson input in the first, and the exponential service time of the second, they also have many special properties. Results can be derived by a variety of quick, clever arguments. Occasionally, these hide the general structural properties accounting for the tractability of these models.

During the 1970s, it became clear that, in both cases, the crucial feature was the structure of the transition probability matrices of their embedded Markov renewal processes. In considerable generality, that structure reflects the fact that both models and their generalizations basically are random walks on a semi-infinite, two-dimensional strip of lattice points.

The embedded Markov chain in models of $M/G/1-$type has the structure

$$P_1 = \begin{array}{c} \mathbf{0} \\ \mathbf{1} \\ \mathbf{2} \\ \mathbf{3} \\ \mathbf{4} \\ \end{array} \begin{pmatrix} B_0 & B_1 & B_2 & B_3 & B_4 & ... \\ C_0 & A_1 & A_2 & A_3 & A_4 & ... \\ 0 & A_0 & A_1 & A_2 & A_3 & ... \\ 0 & 0 & A_0 & A_1 & A_2 & ... \\ 0 & 0 & 0 & A_0 & A_1 & ... \\ ... & ... & ... & ... & & \end{pmatrix},$$

where the symbols $\mathbf{0}$ and $\mathbf{i}$ for $i \geq 1$ stand for sets of m_1 and m states respectively.

The embedded Markov chain in models of $GI/M/1-$type has the structure

$$P_2 = \begin{array}{c} \mathbf{0} \\ \mathbf{1} \\ \mathbf{2} \\ \mathbf{3} \\ \mathbf{4} \\ \end{array} \begin{pmatrix} B_{00} & B_{01} & 0 & 0 & 0 & ... \\ B_{10} & B_{11} & A_0 & 0 & 0 & ... \\ B_{20} & B_{21} & A_1 & A_0 & 0 & ... \\ B_{30} & B_{31} & A_2 & A_1 & A_0 & ... \\ B_{40} & B_{41} & A_3 & A_2 & A_1 & ... \\ ... & ... & ... & ... & & \end{pmatrix}.$$

In both cases, the sequence of matrices $\{A_i\}$ consists of substochastic matrices, whose sum A is stochastic. A particular subclass, common to both types, has a block tridiagonal matrix,

$$P_3 = \begin{array}{c} \mathbf{0} \\ \mathbf{1} \\ \mathbf{2} \\ \mathbf{3} \\ \mathbf{4} \\ \end{array} \begin{pmatrix} B_{00} & B_{01} & 0 & 0 & 0 & ... \\ B_{10} & B_{11} & A_0 & 0 & 0 & ... \\ 0 & B_{21} & A_1 & A_0 & 0 & ... \\ 0 & 0 & A_2 & A_1 & A_0 & ... \\ 0 & 0 & 0 & A_2 & A_1 & ... \\ ... & ... & ... & ... & & \end{pmatrix}.$$

For the continuous-parameter case, these are known as *quasi-birth-and-death* processes or *QBD*s. They arise in a wide variety of practical applications in computer performance and telecommunications modelling.

The theory of the Markov renewal processes of the $M/G/1$ and $GI/M/1$ types is now fully developed. A brief summary cannot do justice to its richness. Its original development is presented in my two books [**?**] and [**?**], but both have been greatly supplemented by new mathematical results, algorithmic developments, and generalizations to structures such as tree-like graphs, and others that have appeared since their publication. A review of these recent developments with an extensive bibliography is given in Neuts [**?**].

We shall cite only one, possibly the most widely known matrix-analytic result, the *matrix-geometric* theorem. If a Markov chain with transition probability matrix P_2 is

positive recurrent - and explicit conditions for that are known - then its invariant probability vector $\mathbf{x}$, partitioned as $\mathbf{x_0}$, $\mathbf{x_1}$, $\mathbf{x_2}$, ... into row vectors of appropriate dimensions, satisfies

$$\mathbf{x_i} = \mathbf{x_1}\mathbf{R^{i-1}}, \text{ for } \mathbf{i} \geq \mathbf{1}, \tag{5}$$

where the $m \times m$ matrix R is the minimal, nonnegative solution of the non-linear equation

$$R = \sum_{i=0}^{\infty} R^i A_i. \tag{6}$$

Once the matrix R is known, the vectors $\mathbf{x_0}$ and $\mathbf{x_1}$ are obtained by solving a system of linear equations. The matrix R has a nice probabilistic interpretation. That has led to several insightful algorithms for its numerical computation as, for example, that for QBDs in Latouche and Ramaswami [**?**].

The inclusion of the labels $M/G/1$ and $GI/M/1$ in the names of these Markov renewal processes refers only to their simplest, elementary cases. Their versatility goes far beyond these classical examples. However, the theory also incorporates generalizations of Poisson arrivals in a natural way. For example, the model $MAP/G/1$ is a single-server queue whose arrivals are described by a Markovian arrival process. For that model, an elegant article by Lucantoni [**?**], discusses the matrix analogues of all formulas known for the classical $M/G/1$ queue and shows how the derivations for that scalar case extend in a natural manner to the generalized model.

4 Current Directions - Open Problems

The overview of the matrix-analytic methods presented so far describes only the barest outlines of a fertile area of research with continued rapid growth. Announcements of new results and articles are found in a Matrix-Analytic Bulletin that I send out by e- mail a few times per year. To subscribe, send me a message at **marcel@sie.arizona.edu**.

The development of efficient algorithms to solve the cited non-linear equation for the matrix R and the related equation

$$G = \sum_{i=0}^{\infty} A_i G^i, \tag{7}$$

which arises in models of $M/G/1-$type, is making major progress. We refer to the recent work of Bini and Meini [**?**], Latouche and Ramaswami [**?**], Meini [**?**], and Akar and Sohraby [**?**].

Structured matrices of either type in which the blocks in the transition probability matrix are themselves infinite matrices, or even general operators, arise in the study of dependent queues. In a major paper, Tweedie [**?**] showed that the matrix- geometric theorem generalizes to the operator case. However, at that level of generality, even the equilibrium condition can no longer be stated in explicit form. Several queueing problems

with deceptively simple statements, such as the two-server shortest queue with Poisson arrivals and exponential service times, can be formulated as *QBD*s with infinite blocks. A subject of current interest is the development of well-justified truncations by which such models can be investigated computationally.

The exploration of descriptors for point processes, in which *MAP*s serve as algorithmically tractable benchmark processes, is highly promising. I have already mentioned the existing literature on this, so let me state here one of my thornier unsolved problems. In local poissonification, the numbers of events in a stationary point process during successive intervals of length a are recorded. If k events occur in such an interval, they are replaced by an equal number of points that are independently and uniformly distributed over that interval. That is done for all intervals. The operation clearly preserves the rate of the process. In repeated local poissonification, that operation is applied repeatedly to the resulting point processes with the understanding that the grid of equidistant points is each time placed in a "random" position, so as to preserve stationarity. It is conjectured that, even with a sequence of window lengths a_i which is bounded away from 0! , the successive poissonifications converge to a Poisson process of the same rate.

Acknowledgements. This research was supported in part by NSF Grant Nr. DMI-9306828. The author thanks the organizers of the Summer Institute 1997 at Sozopol for the cordial hospitality extended to his wife and daughter and himself during their first visit to Bulgaria.

Department of Systems and Industrial Engineering
The University of Arizona
Tucson, Arizona 85721, U.S.A.
e-mail: marcel@tucson.sie.arizona.edu